IMPROVED LANGUAGE IDENTIFICATION USING DEEP BOTTLENECK NETWORK

Yan Song[1], *Ruilian Cui*[1], *Xinhai Hong*[1], *Ian Mcloughlin*[1], *Jiong Shi*[2], *Lirong Dai*[1]

[1]National Engineering Laboratory of Speech and Language Information Processing, USTC
[2]Anhui Post and Telecommunication College

ABSTRACT

Effective representation plays an important role in automatic spoken language identification (LID). Recently, several representations that employ a pre-trained deep neural network (DNN) as the front-end feature extractor, have achieved state-of-the-art performance. However the performance is still far from satisfactory for dialect and short-duration utterance identification tasks, due to the deficiency of existing representations. To address this issue, this paper proposes the improved representations to exploit the information extracted from different layers of the DNN structure. This is conceptually motivated by regarding the DNN as a bridge between low-level acoustic input and high-level phonetic output features. Specifically, we employ deep bottleneck network (DBN), a DNN with an internal bottleneck layer acting as a feature extractor. We extract representations from two layers of this single network, i.e. DBN-TopLayer and DBN-MidLayer. Evaluations on the NIST LRE2009 dataset, as well as the more specific dialect recognition task, show that each representation can achieve an incremental performance gain. Furthermore, a simple fusion of the representations is shown to exceed current state-of-the-art performance.

Index Terms— Language Identification, Deep Neural Network, Bottleneck Feature, Representation Learning

1. INTRODUCTION

Spoken language identification (LID) is the process of determining the language identity of a given utterance. As a branch of audio classification, LID approaches mainly consist of two phases: (1) Front-end feature extraction, which converts a given utterance into a discrete token sequence or a set of continuous-valued feature vectors; (2) Back-end modeling, which constructs the representations for LID.

In phonotactic approaches, such as Phone Recognizer followed by Language Modeling (PRLM) and Phone Recognizer followed by Support Vector Machines (PR-SVMs), the utterances are first tokenized into a sequence of phones using a pre-trained phone recognizer (PR). The phonotactic representations are then constructed using an n-gram model to capture the statistics of phonemic constraints and patterns for each language. A similar concept can be found with acoustic approaches, including Gaussian Mixture Model-Universal Background Model (GMM-UBM) and Gaussian Mixture Model-Support Vector Machine (GMM-SVM) methods, where the short-term spectral features such as MEL-Frequency Cepstral Coefficients (MFCC) and Shifted Delta Cepstrum (SDC) are extracted and modeled using GMM.

It can be seen that effective representation (i.e. feature choice) plays an important role in LID. In recent years, besides the acoustic and phonotactic ones, intensive efforts have studied the effectiveness of representations from other domains, including prosodic and articulatory parameters [1] [2], universal attributes [3] [4], lexical knowledge and so on. Furthermore, with the help of modern machine learning techniques, such as discriminative training [5–7], Factor Analysis (FA) [8], [9] [10] and Total Variability (TV) modeling [11, 12], the effectiveness of representations has been greatly improved, especially for long-duration utterances.

However, performance is still far from satisfactory for highly confusable dialects and short duration utterances. This may be because language information is latent and largely dependent on the statistical distribution of extracted features. For short-duration utterances and for dialects, existing representations are clearly insufficient. They are also susceptible to variations introduced by different speech content, speakers, channels and background-noise.

Deep learning based methods may be helpful to address this issue. In [13], Ganapathy et.al proposed a CNN based method directly trained on the LID dataset. In our previous work [14], we showed that deep bottleneck features (DBF), the output from a constricted internal layer of a structured Deep Bottleneck Network (DBN), can effectively mine the contextual information embedded in speech frames. By representing each utterance as an i-vector, these LID systems were shown to achieve excellent performance in the NIST LRE2009 evaluation.

In the current paper, we extend our previous approach by taking further advantage of the DBN structure. Our motivation is that, if the pre-trained DBN can be considered as a bridge from low-level spectral or acoustic features to high-level phonetic features, then the output from different DBN layers may represent a graded mixture of acoustic and pho-

Thanks to National Nature Science Foundation of China (grant, 61172158) and Chinese Universities Scientific Fund (grant Wk2100060008)

netic information. Exploiting this may be advantageous for LID.

Specifically, with a well-trained DBN structure as front-end feature extractor (shown in Fig.1), we evaluate the effectiveness of representations from the internal bottleneck layer (i.e. DBN-MidLayer) and the topmost layer(i.e.DBN-TopLayer), for dialect and NIST LRE2009 tasks. The contributions of this paper can be stated as follows:

- We propose a novel phonetic utterance representation by averaging the outputs from the topmost layer of the DBN. Like [15, 16], senone posteriors[1] are used as the frame-level features. However, we propose a new Hellinger kernel-based similarity measure between utterances, which we will show can achieve better performance using the same phonetic representation.
- We propose to fuse this representation with our previous DBF-TV representation [14], and will demonstrate that excellent performance can be achieved, especially for dialects and short-duration test conditions.

In summary, the resulting system significantly outperforms the current best-performing method (i.e. that introduced by the authors in [14]) for both dialect recognition and short-duration test conditions on NIST LRE2009.

In the following sections, we first briefly describe the process of language identification using the representation based on DNN in Section 2. We then detail the representations extracted from DNN in Section 3, which is followed by experimental results and analysis in Section 4. Section 5 will conclude this work.

2. SYSTEM DESCRIPTION

As shown in Fig.1, the proposed LID system mainly consists of three parts: 1) DBN structure, 2) Representation based on DBN and 3) Similarity measure.

DBN structure. The DBN structure is the same as the one in our previous work [14]. Different configurations of DBN have been evaluated on the NIST LRE2009 dataset, including dimensions of input and internal BN layer. With the optimal configuration, the resulting LID systems can outperform the original one in [14].

Given the Mandarin Corpus with phone-level labels, DBN training starts with an unsupervised pre-training process, in which a generative deep belief network consisting of stacked Restricted Boltzmann Machines (RBM) is obtained using the method described in [17]. After that, a supervised fine-tuning process is applied to optimize the DBN parameters by minimizing the cross-entropy objective function with a standard error back-propagation (BP) algorithm.

[1] Senones are tied-states within context-dependent phones, which are generally used as the basic units for building word pronunciations in state-of-the-art automatic speech recognition systems

Representation based on DBN. Given an utterance, two types of representation can be extracted, i.e. the DBN-TopLayer based on the output of the topmost layer of the DBN, and the DBN-MidLayer based on the output of the internal BN layer; The DBN-TopLayer is constructed by averaging the frame-level posteriors. Intuitively, DBN-MidLayer can be processed in a similar way to the DBN-TopLayer. However, we found in practice that the average of widely different BN features doesn't tell us much about the content of the underlying utterance, leading to inferior LID performance. Thus, the TV modeling technique is used instead in this work to construct the DBN-MidLayer representation.

Similarity measure. With the DBN-TopLayer and DBN-MidLayer representations, the distance measure should be defined for them respectively. Intuitively, a Euclidean distance can be used due to its simplicity and efficiency. However, the DBN-TopLayer is actually a histogram vector that counts DBN outputs. It is known that using Euclidean distance to compare histograms often yields inferior performance compared to using χ^2 or Hellinger kernel. For DBN-MidLayer representation, conventional cosine distance measure is used as [12].

Given the similarity matrices calculated from DBN-MidLayer and DBN-TopLayer, the SVM classifier can be trained respectively. When given a test utterance, we apply a simple score fusion scheme to make the final decision. We will detail the representation and the corresponding similarity measure in Section 3.

3. REPRESENTATION BASED ON DBN

As mentioned, the DBN structure is pre-trained on a Mandarin Corpus with standard pre-training and fine-tuning processes, which can be regarded conceptually as forming a bridge between low-level acoustic input and high-level phonetic information. In this work, we evaluate the effectiveness of two specific layers of the DBN, i.e. the topmost output layer and the internal BN layer. We first introduce an optimal DBN structure used to extract the frame-level features. Then, the representations based on the outputs of the topmost layer and internal BN layer of the DBN structure, namely DBN-TopLayer and DBN-MidLayer are described respectively, followed by discussion of the similarity metric.

3.1. Optimal configuration of DNN structure

The empirically derived optimal DBN structure has 1 input layer, 5 hidden layers and 1 output layer, configured as $n \times 43-2048-2048-43-2048-2048-6004$. Each frame feature comprises 39-dimensional MFCC+$\triangle$MFCC+$\triangle\triangle$MFCC, and 4-dimensional pitch features corresponding to the static pitch, 1st and 2nd derivatives and voiced speech confidence respectively. The DBN input feature is a concatenation of the n frames centered around the current one. By heuristically

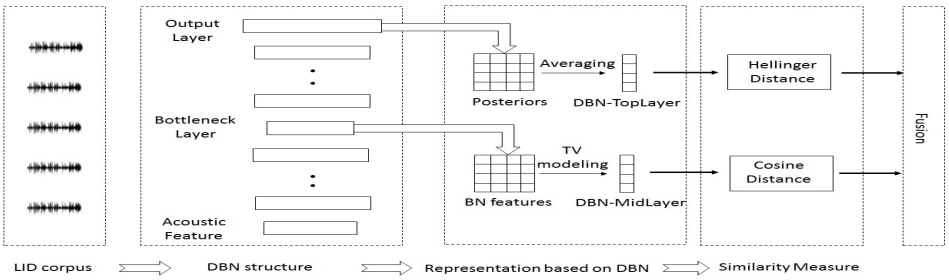

Fig. 1. The LID system using representation based on structured DBN

setting $n = 21$, a 473-dimensional feature vector is obtained. The DBN contains an internal bottleneck layer of 43 hidden nodes, which is much smaller than other layers. This bottleneck layer output forms a compact representation of the input feature, which is considered to be more discriminative and informative than conventional spectral features, i.e. SDC. Similarly to [15, 16], the topmost layer contains nodes corresponding to senones. Conventionally, the senones are automatically defined by a decision tree. For the Mandarin corpus, 6004 senones are used.

3.2. Similarity measure for DBN-TopLayer

Let $\mathcal{Q} = q_k, k = 1, \ldots, K$ be the set of senones. Given a T frame speech utterance, $U = \mathbf{u}_t, t = 1, \ldots, T$, the senone posteriors $p(q_k|\mathbf{u}_t)$ can be predicted by feeding forward the input feature vector $\mathbf{u}_t$ through the DBN structure. As shown in Fig.1, the DBN-TopLayer representation, $[\mathbf{C}_1, \ldots, \mathbf{C}_K]^T$, is the average of the senone posteriors predicted at a frame-level. Each entry in $\mathbf{C}$ can be calculated as

$$\mathbf{C}_k = \frac{1}{T} \sum_{t=1}^{T} p(q_k|\mathbf{u}_t) \tag{1}$$

which is fixed-length feature vector that counts the frequency of senones in the utterance.

Given two utterances $\mathbf{C}_i, \mathbf{C}_j$, the similarity measure $k(\mathbf{C}_i, \mathbf{C}_j)$ defined using a Hellinger kernel is

$$k(\mathbf{C}_i, \mathbf{C}_j) = \sqrt{\mathbf{C}_i^T \mathbf{C}_j} = \sqrt{\mathbf{C}_i}^T \sqrt{\mathbf{C}_j} \tag{2}$$

We can see that calculation of the Hellinger kernel is equivalent to a dot-product of the square-root of the DBN-TopLayer features.

3.3. Similarity measure for DBN-MidLayer

We use the TV modeling technique to extract an i-vector as the DBN-MidLayer representation. Given utterance U, the GMM supervector $\mathbf{M}$ is created by stacking the mean vectors of a GMM adapted to that utterance, modeled as follows

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{3}$$

where $\mathbf{m}$ is the UBM super-vector, $\mathbf{T}$ is a low rank rectangular matrix. $\mathbf{w}$ is the required low-dimensional i-vector with normal distribution $\mathcal{N}(0, \mathbf{I})$. The training process of loading matrix $\mathbf{T}$ is similar to the eigen-voice method [18].

After i-vector extraction, two intersession compensation techniques are applied to remove the nuisance. The first is linear discriminant analysis (LDA) which is a popular dimension reduction method in the machine learning community. Generally, LDA is based on the discriminative criterion that attempts to define new axes minimizing the within-class variance, while maximizing the between-class variance. The second intersession compensation technique we used is within-class covariance normalization (WCCN), which normalizes the cosine kernel between utterances with an inverse of the within-class covariance [11].

If $\mathbf{B}, \mathbf{A}$ denote LDA and WCCN projection matrices respectively, the resulting DBN-MidLayer representation becomes $\hat{\mathbf{w}} = \mathbf{B}^T\mathbf{A}^T\mathbf{w}$. and the cosine distance measure between two utterances $\hat{\mathbf{w}}_i$ and $\hat{\mathbf{w}}_j$ can then be defined as

$$k(\hat{\mathbf{w}}_i, \hat{\mathbf{w}}_j) = \frac{\hat{\mathbf{w}}_i^T \hat{\mathbf{w}}_j}{\| \hat{\mathbf{w}}_i \| \| \mathbf{w}_j \|} \tag{4}$$

4. EXPERIMENTS

To evaluate the effectiveness of the proposed system, we conducted extensive experiments. Firstly on 4 Arabic dialects (Iraqi, Levantine, MSA and Maghrebi) taken from NIST LRE2011, and secondly on the NIST LRE2009 dataset with 23 target languages (Amharic, Bosnian, Cantonese, Creole, Croatian, Dari, English-American, English-Indian, Farsi, French, Georgian, Hausa, Hindi, Korean, Mandarin, Pashto, Portuguese, Russian, Spanish, Turkish, Ukrainian, Urdu and Vietnamese). The training utterances for each language mainly come from two different channels; the dataset of Conversational Telephone Speech (CTS) and the narrow band Voice of America (VOA) radio broadcast dataset.

We have implemented six LID systems based on different representations for evaluation, including phonotactic, acoustic and our proposed DBN-based ones, detailed as follows: **S1:** the PR-SVM system using Russian PR provided by BUT as a front-end feature extractor, with a bag-of-ngram utter-

ance representation, and an SVM classifier trained using the kernel proposed in [19].

S2: a similar PR-SVM system using a Mandarin PR, trained using a front-end DBN feature extractor.

S3: the phonetic system implemented as described in [15,16].

P4: our first proposed system, using the average of senone probabilities taken from the topmost layer of a DBN system, followed by SVM with Hellinger kernel 3.

P5: a better tuned version of the DBN system using TV modeling presented in [14]. In the current implementation, the i-vector dimension is 600. The number of Gaussian components is 512 for the dialect recognition task, and 2048 for the NIST LRE2009 evaluations.

P6: a system which fuses the classification scores obtained from systems P4 and P5.

As illustrated in [20] [21], we use average decision cost function (C_{avg}), and equal error rate (EER) as the performance measurements. The use of these standard evaluation criteria, dataset and evaluation task allow for direct comparison between systems.

4.1. Arabic Dialect Recognition Evaluation

To evaluate the performance of the given systems for dialect recognition, we choose the Arabic task from NIST LRE2011. Results are presented in Table 1, where we can see that the phonetic representation using PR based on a DBN structure outperforms the conventional PR with NN/HMM by about $2\% - 4\%$ (S2 vs. S1). The phonetic system using senone posteriors [15] is better than conventional PR-SVM using a bag-of-trigram representation (the previous best configuration for PR-SVM) (S3 vs. S2). Compared to the senone posteriors [15], an additional absolute 1% performance improvement is achieved by using the proposed DBN-toplayer representation (P4 vs. S3). It can be seen that, for dialect recognition, the DBN-midLayer representation is actually more effective than the phonetic system (P5 vs. P4). This is feasible since the distinction between dialects may be smoothed or degraded by the action of the PR. However, a PR with powerful modeling capability may compensate for this disadvantage. Using frame-level features instead of phones (S3, P4 vs. S1, S2) provides another feasible solution. Furthermore, we can see that the fusion of DBN-TopLayer and DBN-MidLayer achieves the currently best achievable performance. This validates the hypothesis that output from different layers of a single DBN can improve LID performance, by incorporating information from both acoustic and phonotactic representations. Further improvement may be expected by using DBN structure trained with Arabic speech.

4.2. NIST LRE2009 evaluations

To evaluate the performance on a standard evaluation set, we conducted experiments on the LRE2009 dataset. Results are presented in Table 2, where the similar conclusion as dialects can be observed: Namely that the proposed phonetic system using DBN-TopLayer representation (P4) significantly outperforms conventional PR-SVMs (S1, S2) and the system using senone posteriors (S3). The tuned DBN-MidLayer (P5) using the configuration as illustrated in Section 3 achieved much better performance than either the previously reported one [14] or the best reported results from NIST LRE 2009 [12] (both results are listed at the end of the table). Again, P6, the fusion of both acoustic and phonetic systems, performs best overall, by a significant margin.

Table 1. Evaluations on Arabic recognition in terms of EER and C_{avg} (%)

System	30s	10s	3s
S1	7.93/7.69	17.27/16.95	30.13/29.57
S2	5.16/4.96	13.40/13.10	28.71/28.33
S3	4.13/4.06	12.48/12.24	26.19/25.98
P4	3.92/3.84	11.49/11.31	25.02/24.36
P5	2.70/2.56	7.63/7.17	19.47/19.03
P6	**2.34/2.26**	**7.13/6.94**	**18.48/18.40**

Table 2. Evaluations on NIST LRE2009 in terms of EER and C_{avg} (%)

System	30s	10s	3s
S1	2.58/2.32	7.29/7.21	21.41/21.67
S2	2.08/3.03	6.79/6.84	20.93/21.56
S3	1.56/1.53	4.34/4.30	16.67/16.57
P4	1.54/1.52	3.78/3.78	14.28/14.23
P5	1.32/1.29	2.52/2.60	9.84/9.84
P6	**1.20/1.16**	**2.40/2.38**	**8.95/8.91**
DBF-TV [14]	1.98/1.97	3.47/3.45	9.71/9.74
SDC-TV [12]	2.40/	4.80/	14.20/

5. CONCLUSION

This paper has built on the previous work of the authors which demonstrated state-of-the-art performance on the NIST LRE2009 LID task using features extracted from a deep bottleneck network (DBN). Motivated by the observation that the deep neural network acts as a bridge spanning between a purely acoustic feature input and a purely phonotactic classification, this paper proposed using a fusion of representations extracted from a single well-trained DBN. It is well known that both acoustic and phonotactic features can be applied to the LID task, and that each have their own strengths. Therefore this paper proposed and explored fusing these strengths. Results exhibit excellent performance for both dialect recognition and LID, on NIST LRE tasks. In addition, this paper proposed using a novel averaging method, with Hellinger kernel based similarity measure, for the top level DBN posteriors, which was shown to perform well.

6. REFERENCES

[1] Driss Matrouf, Martine Adda-Decker, Lori Lamel, and Jean-Luc Gauvain, "Language identification incorporating lexical information.," in *Proc. of ICSLP*, 1998, vol. 98, pp. 181–184.

[2] Stephen J Eady, "Differences in the F0 patterns of speech: Tone language versus stress language," *Language and Speech*, vol. 25, no. 1, pp. 29–42, 1982.

[3] Sabato Marco Siniscalchi, Jeremy Reed, Torbjørn Svendsen, and Chin-Hui Lee, "Exploring universal attribute characterization of spoken languages for spoken language recognition.," in *Proc. of InterSpeech*, 2009, pp. 168–171.

[4] Hamid Behravan, Ville Hautamaki, Sabato Marco Siniscalchi, Elie Khoury, Tommi Kurki, Tomi Kinnunen, and Chin-Hui Lee, "Dialect levelling in finnish: A universal speech attribute approach," in *Proceedings of Interspeech 2014*, 2014.

[5] Dan Qu and Bingxi Wang, "Discriminative training of GMM for language identification," in *Proc. of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[6] Lukas Burget, Pavel Matejka, and Jan Cernocky, "Discriminative training techniques for acoustic language identification," in *Proc. of ICASSP*, 2006, vol. 1, pp. 209–212.

[7] Fabio Castaldo, Daniele Colibro, Emanuele Dalmasso, Pietro Laface, and Claudio Vair, "Acoustic language identification using fast discriminative training.," in *Proc. of InterSpeech*, 2007, vol. 7, pp. 346–349.

[8] Claudio Vair, Daniele Colibro, Fabio Castaldo, Emanuele Dalmasso, and Pietro Laface, "Channel factors compensation in model and feature domain for speaker recognition," in *Proc. of Odyssey: Speaker and Language Recognition Workshop*, 2006, pp. 1–6.

[9] Fabio Castaldo, Daniele Colibro, Emanuele Dalmasso, Pietro Laface, and Claudio Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Trans Audio Speech Lang Processing*, vol. 15, no. 7, pp. 1969–1978, 2007.

[10] Valiantsina Hubeika, Lukas Burget, Pavel Matejka, and Petr Schwarz, "Discriminative training and channel compensation for acoustic language recognition.," in *Proc. of InterSpeech*, 2008, pp. 301–304.

[11] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans Audio Speech Lang Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[12] Najim Dehak, Pedro A Torres-Carrasquillo, Douglas A Reynolds, and Reda Dehak, "Language recognition via i-vectors and dimensionality reduction.," in *Proc. of InterSpeech*, 2011, pp. 857–860.

[13] S. Ganapathy, K. J. Han, S. Thomas, M. Omar, M. V. Segbroeck, and S. Narayanan, "Robust language identification using convolutional neural networks," in *Proceedings of Interspeech 2014*, 2014.

[14] Yan Song, Bing Jiang, YeBo Bao, Si Wei, and Li-Rong Dai, "I-vector representation based on bottleneck features for language identification," *Electron Lett*, vol. 49, no. 24, pp. 1569–1570, 2013.

[15] Luciana Ferrer, Yun Lei, Mitchell Mclaren, and Nicolas Scheffer, "Spoken language recognition based on senone posteriors.," in *Proceedings of the Interspeech 2014*, 2014.

[16] Yun Lei, Luciana Ferrer, Mitchell Mclaren, and Nicolas Scheffer, "Application of convolutional neural networks to language identification in noisy conditions," in *Proceedings of Odyssey 2014*, 2014.

[17] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[18] Patrick Kenny, Gilles Boulianne, and Pierre Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans Speech Audio Process*, vol. 13, no. 3, pp. 345–354, 2005.

[19] P.A.Torres-Carrasquillo and et.al. E. Singer, "Discriminative training techniques for acoustic language identification," in *Proceedings of ICASSP 2006*, 2006, pp. 209–212.

[20] Alvin Martin and Craig Greenberg, "The 2009 NIST language recognition evaluation," in *Proceedings of Odyssey 2009: The Speaker and Language Recognition Workshop*, 2010, pp. 165–171.

[21] Alvin Martin, Craig Greenberg, M. Howard Hohn, George R.Doddingtong, and John J. Godfrey, "Nist language recognition evaluation-past and future," in *Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014.